DOCUMENT RESUME

ED 459 212 TM 033 511

AUTHOR Loomis, Susan Cooper
TITLE Judging Evidence of the Validity of the National Assessment of Educational Progress Achievement Levels.
INSTITUTION ACT, Inc., Iowa City, IA.
SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE 2001-06-00
NOTE 40p.; Paper presented at the Annual Meeting of the Council of Chief State School Officers (Houston, TX, June 24-27, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; *Cutting Scores; Elementary Secondary Education; *Evaluation Methods; National Surveys; Standards; *Test Results; *Validity
IDENTIFIERS Competency Tests; *National Assessment of Educational Progress; *Standard Setting

ABSTRACT

This paper describes (1) the procedures developed to set achievement levels for the National Assessment of Educational Progress (NAEP) that contribute to establishing the validity of the levels and (2) the research studies designed to collect information related to the validity of the achievement levels and the outcomes of the process. The central issue in examining the validity of standards is whether there is evidence of procedural validity. The standards must be generally accepted as reasonable for the outcomes of the process for setting cutpoints to be valid. For each of the three American College Testing (ACT, Inc.) program contracts with the National Assessment Governing Board, the process of developing achievement levels descriptions has been different, as described, but in all cases there has been an effort to solicit broad-based commentary about the reasonableness of the achievement level descriptions. The selection of the panelists is important, since standard setting panels must be seen as credible. The paper describes the selection of panelists, field trials and pilot studies, training for facilitators, and panelist training. Several different rating methodologies have been evaluated and tested in panel studies for the NAEP achievement level process, but the modified Angoff method has the most solid research base in standard setting. Panelists participate in three rounds of item-by-item ratings with a variety of feedback after each round completing evaluations throughout the process. ACT, Inc. has performed various types of evaluations of the standard setting process and data. These include analyses of standard-setting data that are somewhat standard and research studies related to validation that are further divided into studies using item mapping procedures, studies comparing teachers' judgments of performance to empirical classifications of student performance, and studies comparing judgments of performance represented in test booklets to the empirical classification of these booklets. In the end, there is no way to know with certainty that cutscores are valid, although substantial effort goes into ensuring procedural validity. (Contains 21 tables and 50 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 459 212

Judging Evidence of the Validity of the National Assessment of Educational Progress Achievement Levels

by

Susan Cooper Loomis
Senior Research Associate
Policy Research Department

ACT, Inc.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

S.C. Loomis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not neces sarily represent official OERI position r r policy.

TM033511

This paper was prepared for presentation at the CCSSO Large-Scale Assessment Conference, June 24-27, 2001, Houston.

The research reported in this paper was supported by contracts with the National Assessment Governing Board.

2 BEST COPY AVAILABLE

Judging Evidence of the Validity of the National Assessment of Educational Progress Achievement Levels

Susan Cooper Loomis
Senior Research Associate
Policy Research Department
ACT, Inc.

Introduction

When training panelists in the standard-setting process and informing them about the purposes of setting standards, one is likely to talk about how standard setting helps to answer the question: *How much is enough*? That's a great question to be able to answer, and panelists seem immediately to perceive the benefits of setting standards to answer it.

A slightly different question is often asked about the outcomes of a standard-setting procedure. It is also a great question, but a great answer is not so easy to find. How can you know that students *can* do the things that the standards say they *should* be able to do? This question focuses attention on the issue of the validity of the standards.

The procedures developed to set achievement levels for the National Assessment of Educational Progress (NAEP) that contribute to establishing the validity of the levels will be presented in this paper. In addition, the research studies designed to collect information related to the validity of the achievement levels and the outcomes of the process will be described.

The procedures described here were developed under three different contracts awarded to ACT by the National Assessment Governing Board (NAGB) from 1991 and 2001. These procedures were developed with the advice of numerous experts. Members of the Technical Advisory Committee on Standard Setting are the principal contributors.[1]

Procedural Validity

Perhaps the central issue in examining the validity of standards is whether there is evidence of *procedural validity*. Kane (2001) points out that "procedural evidence is often considered adequate to provide basic support for the performance standards and cutscores unless there is conflicting evidence suggesting that the performance standard or cutscore is inappropriate." (p. 64) He goes on to point out that a standard-setting process would not be judged to be valid if there were a lack of evidence of procedural validity, but evidence of procedural validity does not assure the validity of the process. Procedural validity is a necessary—but not a sufficient—condition for validity.

There are many different sources of guidelines and lists of good practices associated with setting standards and providing evidence of the validity of the outcomes. Cizek, 1996; Hambleton, 2001;

---

[1] Members of the Technical Advisory Committee on Standard Setting (TACSS) include William Brown, Barbara Dodd, Robert Forsyth, Ronald Hambleton, John Mazzeo, William Mehrens, Jeff Nellhaus, Mark Reckase, Douglas Rindone, Wim van der Linden, and Rebecca Zwick. Robert Brennan has served on the ACT Technical Advisory Team (TAT) and as a representative of that team to the TACSS for the entire period. Reckase also served on TAT before joining the TACSS. Forsyth, Hambleton, Mehrens, Reckase, and Brennan have been technical advisors for the NAEP Achievement Levels-Setting (ALS) contract since 1991. Michael Kane, Brenda Loyd (deceased), Eugene Johnson, and James Carlson were all former members of TACSS.

Kane, 1994; 1995; 2001; Mehrens, 1995; and Mehrens and Cizek, 2001 are but a few. Each of those authors has contributed significantly to the NAEP ALS process, and much of what is presented here will have been shaped by their input over the years. Rather than focusing on a single list, however, this paper will survey relevant procedures and attempt to highlight ways in which each can be used in establishing the validity of the NAEP ALS Process.

Statements of the Standards

The statement of the standards must be generally accepted as reasonable in order for the outcomes of the process for setting cutpoints to be valid. If the statements of what students should know and be able to do are not judged to be reasonable, the cutpoints cannot be judged to be reasonable. Thus, having general agreement on the reasonableness of the statements of standards is a necessary condition for valid standards.

The NAEP Achievement Levels Descriptions are the *statements of the standards* for a subject and grade. NAGB determined that there would be three achievement levels or goals for NAEP: Basic, Proficient, and Advanced.[2] Further, NAGB defined each goal in general terms, and those are referred to as the *policy* definitions. The policy definitions are general statements describing student performance at each level. There is no mention of grade level and no mention of subject matter. These policy definitions serve as the general calibrators for the achievement levels descriptions. The achievement levels descriptions are operational definitions of the policy definitions, taken from the framework, to describe what student should know and be able to do at each achievement level for each grade.

For each of ACT's three contracts with NAGB, the process of developing achievement levels descriptions has been different. In all cases, there has been a genuine effort to solicit broad-based commentary regarding the reasonableness of the achievement levels descriptions. The differences have been with respect to when the achievement levels descriptions were developed and when reviews have occurred.

In the 1992 NAEP ALS processes implemented to set achievement levels in mathematics, writing, and reading, panelists were asked to develop statements of what students should know and be able to do in order to meet the general definitions of the three performance levels (Basic, Proficient, and Advanced) given by NAGB. Working with the policy definitions, panelists engaged in brainstorming sessions to develop operational definitions of performance at each level for the subject and grade level of the panel. Those definitions were then evaluated and modified by panels of teachers and curriculum specialists who participated in validation studies for each subject. The Achievement Levels Definitions (ALDs) were shared with various stakeholder groups in a series of public comment forums to collect additional input and recommendations before adoption by NAGB.

Preliminary achievement level definitions were included in the development of the subject frameworks for geography, U.S. history, and science—the subjects for which achievement levels were set in ACT's second contract with NAGB, 1993–1997. The preliminary ALDs were reviewed as part of the framework development review, and that generally includes several large-scale public comment forums scheduled throughout the nation. During the ALS processes in 1994 and 1996, panelists evaluated the preliminary achievement levels definitions and modified them if necessary. In fact, few changes were made to any of the preliminary definitions. Following the ALS process, the achievement levels descriptions were included in review packets distributed as part of the collection of public commentary on the achievement levels prior to adoption by

[2] See Public Law 100-297 (1988).

NAGB.[3] Public comment was collected in meetings scheduled in Washington, D.C. and by letters from individuals and representatives of groups, agencies, and organizations.

The format for the first two contracts had been to have the development or modification of ALDs by panelists *during* the ALS process and to have the large-scale review *after* the cutpoints were set. ACT's 1998 proposal included a plan for finalizing the achievement levels definitions *prior* to convening the achievement levels-setting (ALS) panels. Review panels were convened throughout the country, and the members of the review panels were selected according to the guidelines used for identifying ALS panelists. Their recommendations were collected for evaluation and implementation by a small panel of content experts who had developed the subject frameworks for NAEP. The review process was iterative. ACT collected public commentary from stakeholder groups for additional review and evaluation by the content experts. Modifications and reviews continued until general consensus seemed to have been reached. The process of finalizing the ALDs was comprehensive and thorough.[4]

There was some concern among ACT staff and TACSS members that panelists would not have the same "buy-in" to those definitions—and the process—as they had in previous procedures when the panelists were allowed to shape the statements of the standards. There was no evidence of this, however. Panelists reported levels of satisfaction with the ALDs and levels of understanding their meaning that were equal to those reported by panelists who spent hours modifying the definitions (Loomis & Hanick 2000b; 2000c). The 1998 panelists had more time to spend on forming a clear concept of borderline performance and on other aspects of the process because they were not involved in modifying the definitions. This plan worked very well.

Having the statements of standards set before the standard-setting panels are convened seems to be a worthy goal. If the review process is held *after* the standard-setting panels are convened, then the policy board is responsible for making any adjustments in the cutscores that may seem appropriate in light of recommendations collected in the review. If substantive changes to the statements are recommended, then the cutpoints may no longer serve as translations of the statements to the score scale. In that case, it would be necessary to either change the relationship among the framework, statements of standards, and cutpoints or to convene (reconvene) another panel of standard setters. Having a thorough review of the statements of the standards prior to setting the cutpoints seems the best alternative by far.

Panelist Selection

It seems unlikely that education standards will be viewed as credible if the standard-setting panels are not viewed as credible. Panelists who set standards for public education must be broadly representative. Panelists who set standards for public education must be qualified. They must understand student behavior and have some knowledge of the knowledge and skills required of students in the grade level for which they serve as panelists. These two criteria—broadly representative and well qualified—create a challenge for forming standard-setting panels.

---

[3] The 1996 Science achievement levels developed through the ACT-NAGB ALS Process were not adopted by NAGB. NAGB judged that the outcomes of the process were not reasonable, and they decided on different cutpoints for most levels: some higher; some lower. As a result of the changes, new descriptions had to be developed for reporting student performance relative to the Science NAEP cutpoints. The Science NAEP ALDs are based on the items for which students scoring within the cutpoint ranges had at least a 65% average probability of correct response. The descriptions do not necessarily reflect the NAGB policy definitions of Basic, Proficient, and Advanced performance for each grade in science.

[4] The process is described and completely documented in Loomis and Hanick, 2000.

NAGB decided from the beginning that the NAEP ALS process should include representatives of the general public as well as educators. They specified that approximately 55% of the panelists be classroom teachers in the subject and grade levels for which achievement levels were set; 15% other educators—counselors, curriculum directors, higher education faculty in the subject, and so forth; and 30% general public. ACT developed guidelines for panelists of each type to help assure that there was a reasonable expectation that panelists were qualified to make judgements about students in both the subject matter and the grade level for which they would serve as panelists.

The process of identifying and selecting NAEP ALS panelists has been well documented and reviewed. (See Raymond & Reid, 2001.) ACT included a complete description of the plan in the Design Document for each of the ALS procedures, and the documents were sent out for public review and commentary by stakeholder groups prior to implementation in the first two contracts.[5]

The plan incorporates principals of statistical sampling procedures. A national database of school districts serves as the primary sampling unit. Nominators are identified in each district, and they are invited to submit names of persons who meet the guidelines (distributed to nominators) for panelist selection. Nominators must supply information about the candidates that is then used for purposes of selecting panel members.

The most well qualified nominees are given first priority for selection. In addition, panels are drawn to be representative with respect to panelist type (teachers, other educators, and general public—according to NAGB percentage requirements), gender (as nearly equal as possible), region (as nearly equal as possible), and race/ethnicity (as nearly proportional to the U.S. population as possible). The specific features to be represented on the panels changed somewhat over time. Initially, region was a factor included in drawing the sample of districts to identify nominators, but it was not a factor for equal representation on the panels in the first contract period. The aim for representation by race/ethnicity was to have at least 20% of the panelists from minority groups, in general. This changed in the 1998 process so that proportional representation by specific racial/ethnic groups (Asian Americans, African Americans, Hispanic and Mexican Americans, and Native Americans) was the goal.

Twenty panelists were recruited for the ALS panels for each subject and grade level in the 1992 process. The number of panelists was increased to 30 for each in subsequent ALS procedures. Pilot study panelists were selected through exactly the same procedures used for the ALS process, and pilot studies generally included 20 panelists for each grade in a subject. Panel sizes for field trials and validation research studies varied according to the design requirements. As a general rule, the aim was to have at least 10 panelists participating in a procedure or research group.

The method of selecting panelists was submitted to broad-based review and evaluation. The process has generally been found to be thorough and effective as a means for selecting representative panels of qualified panelists that meet the specified requirements established by NAGB and recommended by various stakeholders and the TACSS.

---

[5] The stakeholder lists were modified for each subject to include key agencies and organizations, but many umbrella-like education agencies and organizations were included on each list of approximately 200 individuals and groups.

Training

Field Trials and Pilot Studies

One full-scale pilot study is now conducted for each subject before the operational ALS is implemented. There was only one pilot study for the 1992 ALS cycle. Panelists were selected from the St. Louis area, and the selection criteria now in place were not used.

Twenty panelists were recruited for each pilot studies in geography and U.S. history for the 1994 ALS cycle. Those pilot studies were used for collecting research data on several alternative methods and procedures. ACT staff recommended against that plan for the next cycle.

Because the Science NAEP included hands-on tasks, two pilot studies were conducted. The first study was to train in the use of hands-on materials, to study the effects of some unusual scoring procedures for some constructed response items in science, and to get a sense of the amount of time needed to work with the science assessment. The first study included 10 panelists at grade 8, 20 at grade 4, and 30 at grade 12. The second pilot study was a full-scale test of procedures planned for the operational ALS. One procedure to collect information on item mapping criteria was included.

In 1998, a series of research studies were proposed to prepare for the operational ALS process. Five different panel meetings were convened before the pilot studies. These field trials were to test out procedures and collect research information related to procedures planned for the 1998 ALS cycle. The pilot study in each subject was a full-scale practice run for the ALS. A de-briefing session was also held after each pilot study with a representative sample of panelists at each grade level. A prepared list of discussion topics was distributed, and panelists were urged to comment on any additional matters that needed to be discussed. Panelists were aware that there would be a de-briefing, and they often shared their comments with persons who had been selected to participate in the de-briefing. Several changes in the agenda were made, and some procedures were added or modified as a result of the de-briefings.

Facilitators

Training is an issue for facilitators, as well as for panelists. Process facilitators are trained extensively in a series of half-day and whole-day sessions over a period of several weeks. Training continues until they are thoroughly familiar with the procedures and with the content of the materials used in each session. Starting with the 1994 ALS cycle, detailed outlines have been provided to the facilitators that describe each step in the process. In addition, instructional/training materials presented in the general session are copied and distributed to each grade facilitator so that reviews of procedures and instructions are consistent across the grade groups. Every effort is made to assure consistency in training across grade groups. As the process evolved, it became apparent that process facilitators must be well trained in quantitative analysis—ideally, in measurement. Process facilitators must agree to be team players, and they must possess a high level of *people* skills.

Changes in the content staff were made over time as well. Content facilitators in the first contract were generally highly experienced ACT staff working in test development in the specific subject area. Key persons from the framework committees in each of those subjects worked with the ACT staff to prepare for the task of training panelists in the framework and developing achievement levels descriptions. Starting with the second contract, content facilitators have been selected from among the persons who served on the framework committees. They have first-hand information about decisions regarding the content, format, and organization of the assessment and about the development of the achievement levels descriptions. Recommendations from persons who worked with these committees are helpful in selecting facilitators who are likely to work

well in the standard-setting context. Most of the content facilitators have been involved in all stages of the development of the NAEP in a subject before participating in the ALS process.

A full-day training session including both content and process facilitators is held prior to the pilot study. Two-three hour review/retraining sessions are held before the start of the panel meetings—both pilot studies and ALS meetings. With only one or two exceptions all facilitators participated in the pilot study before the ALS meetings.

It is not to monitor activities simultaneously in all grade groups, and it is not desirable to need to do so. Selecting the right people to serve as facilitators is very important and providing them with the necessary training and instructional support is essential to a successful standard-setting process. Panelists must feel confident that the facilitators are entirely competent and well trained in the process and tasks.

Panelists

Standard-setting panelists must be well trained. They must be trained to understand the purposes of setting standards, the assessment and scoring protocols, the statements of standards, the rating method(s), and the feedback. The credibility and defensibility of the process and the outcomes of the standard-setting process are a function of the level of agreement among panelists on the meaning of the statements of the standards and on the performance required of borderline students. The credibility and defensibility of the process are also a function of the extent to which panelists appear to understand the tasks and to feel confident in and satisfied with their ability to perform the tasks. This requires training for panelists—especially broadly representative panelists.

The NAEP ALS Process probably provides the most extensive and intensive training for panelists of any standard-setting process. A detailed description of the procedures for training panelists can be found in Loomis and Hanick (2000b; 2000c). Raymond and Reid (2001) provide a rather detailed description and discussion of the NAEP training procedures, and they give the procedures a favorable review. In part, the extensive and intensive training is required by the combination of types of panelists included in the process.

Training begins early—shortly after panelists are selected. Three sets of advance materials are sent to panelists, along with detailed letters instructing them in the use of the materials and preparing them for the process. Advance mailings are spaced at approximately 10-14 day intervals.

Once on site, panelists participate in instructional sessions and hands-on training sessions, and they do not start rating items until the afternoon of the third day of the process. During the training period, panelists are first given a comprehensive overview of the process and purposes for setting NAEP Achievement Levels. The general orientation and overviews are presented by both the ACT general facilitator and by the NAGB staff person in charge of achievement levels. Panelists take the NAEP under timed conditions and review their work with answer keys and scoring rubrics. They spend most of the first 2½ days gaining an understanding of the meaning of the achievement levels descriptions and forming a clear concept of borderline performance. In addition, panelists become familiar with the assessment framework, the item pool, the scoring rubrics, and other key features of the assessment. Because most NAEP item pools include a large number of constructed response items, a large amount of time is needed to help panelists become familiar with the scoring rubrics and to have a clear, realistic understanding of student responses to open-ended questions. Panelists are given ample opportunity for discussions with other

panelists in their subject group and for asking questions and seeking clarification from process and content facilitators.

All instructions and initial training sessions take place in general sessions including all panelists from each grade level. This assures that every panelist hears the same instructions and receives the same training in the procedures. Procedures are implemented in grade-group sessions led by process facilitators. The amount of time required for each procedure is carefully estimated on the basis of data collected through field trials, pilot studies, and previous operational ALS studies. Too little time to complete a task adds to the frustration of panelists and jeopardizes the possibility of a valid outcome. Similarly, too much time is likely to lead panelists to redefine the task and to have an unintended result. Thus, accurate time estimates are very important to the substantive outcome of the process, as well as to the logistic aspects of the process.

Panelists participate in training that focuses their attention on the achievement levels descriptions and borderline performance. Training includes exercises at the individual item level and holistically, over a NAEP test booklet. Training materials in exercises do not include items that will later be rated by panelists. This means that judgements made in this training period will not influence judgements to be made after training is completed and ratings are underway. At the same time, however, the training materials allow panelists to become familiar with all the items in the NAEP for their grade level. Once panelists are trained for the rating task, they participate in the first round of ratings using only the achievement levels descriptions, concepts of borderline performance or written borderline descriptors, and the information gathered in training. Once the item-by-item rating sessions begin, discussions among panelists are not allowed.

Rating Methodology

Several different rating methods have been evaluated and tested in panel studies for the NAEP ALS Process. The modified Angoff method, and variants of it, has been used for setting all NAEP achievement levels. Because the method for collecting judgments is frequently at the center of standard-setting decisions, and because the method used for the NAEP ALS Process has been the criticized (National Academy of Education, 1993; National Research Council, 1999), a description of several methods tested by ACT is presented here, along with the general rationale for the acceptance or rejection of the method(s). A more complete review of the various methods is presented in Loomis and Bourque, 2001.

For ACT's first standard-setting contract, NAGB specified that the modified Angoff method be used. In addition, ACT designed a Paper Selection method for use with the constructed response items in the 1992 ALS process. The choice of methods was left to ACT in the second contract awarded, although NAGB specified that the method selected should have a sound research base and should not be likely to produce achievement levels that were greatly different from those already set in 1992.

ACT tested several variants of the modified Angoff method in pilot studies for the 1994 ALS process. The final decision was to use the modified Angoff method for multiple-choice items and a mean estimation procedure—quite similar to the modified Angoff method—for constructed response items. The Technical Advisory Committee on Standard Setting recommended that a paper selection procedure be implemented prior to the rating process as part of training in borderline performance and preparation for rating constructed response items. This was, in part, recommended because of the research base established for the paper selection procedure in the 1992 ALS process.

In choosing the rating methodology, TACSS has counseled that the rating method used to establish cutpoints on the score scale should be compatible with the scaling method used to put student scores on the reporting scale. Kane (2001) suggests that the method "be consistent with the design of the assessment procedure, and both the standard-setting method and the assessment procedure should be consistent with the conception of achievement underlying the decision process." (p.59) For example, a noncompensatory rating method should not be used to set cutpoints for an assessment using a compensatory scaling model.

Research by ACT has consistently revealed that panelists tend to assign more weight to items requiring constructed responses than to multiple-choice items (Loomis & Hanick, 2000b; 2000c). Panelists also tend to select some items that they consider essential indicators of performance at a particular level of achievement and to weight them disproportionately when judging student performance holistically. So, no mater how well the student performs on the overall assessment, the panelist may perceive that the student has a relatively low level of achievement due to the failure to correctly answer a particular question or set of questions related to a particular area of knowledge or skill. Given the fact that NAEP uses a compensatory scaling model, an item-by-item rating method has been favored over holistic methods for the NAEP ALS Process.

The cutpoints should result from the judgments of panelists and not from the methodology. ACT experimented with the use of item mapping procedures, similar to the Bookmark methodology. One proposed plan was to have panelists use an item-by-item rating process for two (or three) rounds and then switch to the item maps for selecting the final cutpoints. This combination of methods was tested in a field trial for civics and found to work well (Loomis, Bay, Yang & Hanick, 1999; Loomis, Hanick, Bay, & Crouse, 2000a). The need to have a probability value to map the items to the scale, however, was judged to be a problem with this method. See Loomis and Bourque (2001) for an example and discussion of this problem. A lower probability used in mapping items would result in a relatively lower cutpoint and a higher probability value would result in a higher cutpoint.

Once the panelists have made their judgments about where to draw the boundary between two achievement levels (where to set the cutpoint), the actual value of the cutpoint on the score scale must be determined. That cannot be done without some decision regarding the p-value for mapping items. There was no established criterion for determining how to map (locate) items on the score scale. NAGB had established no such criterion and none existed in the research literature. Indeed, the research conducted to date has revealed no consensus on the choice of p-values for mapping.[6] In the absence of a policy decision to determine the p-value, TACSS recommended against further research regarding the use of the item mapping method.

ACT tested a variant of the modified Angoff method—the *yes/no* method or *correct/incorrect* method of rating items in the first set of field trials for the 1998 NAEP ALS Process. Impara and Plake (1996) had reported that panelists found the method easier to use than the modified Angoff procedure that required estimates of percentages of borderline students who would correctly answer questions. Panelists who used the method in the field trials conducted by ACT reported that it was easy to use and conceptually clear, but the responses of those field trial panelists were no less positive than the panelists who had estimated percentages in other ALS studies (Loomis, Bay, Yang, and Hanick, 1999; and Loomis, Hanick, Bay, & Crouse, 2000a; 2000b). Further analyses by ACT (Reckase, 1998; Reckase & Bay, 1999) revealed that the Item Score String

[6] ACT conducted several research studies as part of the 1996 Science NAEP ALS Process (ACT, 1997). Zwick, et. al (2000) examined this issue with respect to the choice of p-value to use in selecting exemplar items. Kolstad has researched the issue extensively as well. See Kolstad (1996) and Kolstad, et. al (1998).

Estimation method, ACT's name for the method, was biased. The cutpoint for the lower level would be set lower than the panelist intended—lower than the true score estimates of borderline performance, and the cutpoint for the higher level would be set higher. No further research was conducted on this method.

ACT also tested the use of a booklet classification method for the 1998 Writing ALS. ACT had used a booklet classification process in several validation studies (ACT, 1995c; 1997), and the results had always indicated that the cutpoints set using the modified Angoff/mean estimation methods would need to be set higher to match the performances required by the booklet classification method. The method had not been used in a standard-setting context, however, until the 1998 field trials for writing. The cutpoints were not really higher than the item-by-item method *in this particular instance*, but this was the exception to findings from ACT research with the method. Another consideration was the lack of a reliable method for computing cutpoints, and this was the ultimate reason for eliminating the booklet classification method from further consideration. Several different methods for computing cutpoints were tested, and each resulted in different cutpoints. (See Hanson, 1998; Hanson, Bay, & Loomis, 1998; Loomis, Hanick, Bay & Crouse, 2000b.) Further, the logistic requirements for implementing a booklet classification process are great. Choices must be made about the number of booklets to include in the classification, the number of different test booklet forms (items) to include, and the distribution of performances in the booklets. (See Bay, 1998.) Concerns related to copying, transporting, and handling the huge volume of secure test materials represent a further challenge to the choice of using a booklet classification procedure. These factors must all be taken into account when selecting a rating method to use, and the costs of the booklet classification method turn out to be quite high.

ACT used the last set of field trials in civics and writing in 1998 to test a new procedure designed by Mark Reckase. The lack of a solid research base caused concern regarding the use of the Reckase Methods in the ALS process. ACT and TACSS felt the charts added greatly to the understanding of panelists, but we were not sure about the method, per se, of setting cutpoints. While the Reckase Method was not implemented after the field trials, the central feature of the method—the Reckase Chart—was incorporated into the 1998 NAEP ALS Process.

In summary, the modified Angoff method has the most solid research base of all standard-setting methods (Cizek, 1993; Mehrens, 1995). ACT has directly collected research data through field trials and pilot studies for all procedures implemented in the NAEP ALS Process. Loomis and Bourque (2001) suggest six criteria for the choice of methods to use in the NAEP ALS Process. (1) Panelists generally report that the rating method is conceptually clear and easy to use.[7] It is simple to explain. (2) The method for collecting item ratings to set cutpoints is consistent with the method of collecting student performances and reporting them on the score scale. (3) The method allows panelists to consider scoring conventions (how missing data are handled in NAEP) and various aspects of student test-taking behavior without having these considerations impact their item ratings and resulting cutpoints. (4) The method for computing cutpoints is consistent with the scaling model and results in unbiased representations of panelists' ratings. (5) The method maximizes the use of available data and minimizes the loss of data (a perceived flaw of some variants of the modified Angoff tested in 1994 pilot studies). (6) Policy decisions are left to policymakers.

---

[7] For ACT research, please see ACT 1993a – 1993c; 1994a; 1994b; 1995b; 1997a; Loomis & Bourque, 2001; Loomis & Hanick, 2000b; 2000c. For further confirmation of this point, see Mehrens, 2001.

Rating in Rounds with Feedback

Panelists participate in three rounds of item-by-item ratings with a variety of feedback following each round. Panelists rate each item at all three NAEP achievement levels before rating the next item. This reduces the amount of time needed in the rating process, and it reduced the probability of "illogical" ratings whereby the percentage estimate recorded for a higher level of performance is actually lower than that for a lower level. Panelists are allowed to choose the order for rating levels for each item so they may rate an item in any order, with respect to the levels.

All panelists rate all items in their item rating pool. The item rating pool for each grade is divided so approximately half of the items are rated by panelists in each of two item rating groups. Item rating pools and item rating groups are divided so that each half is statistically equivalent to the other—to the maximum extent possible. Each panelist actually rates more than half of the item in the item pool for the grade because at least one set of items in a NAEP block (a section in the test booklet) is assigned as a common block to the item rating pool of all panelists in the grade group. This set of common items allows comparisons of ratings by panelists in each rating group to determine whether the cutpoints that would be set by the two groups of panelists on the same set of items are statistically different. This feature of the ALS process—a split-half design—for the item rating process, allows ACT to analyze the data for each grade group as a quasi-replication study. Findings from analyses of ratings by the two rating groups show that the differences are generally not statistically significant (Loomis & Hanick, 2000b; 2000c).

Feedback data are provided to panelists after each round of item–by–item ratings. For more detailed information about the feedback data used in the NAEP ALS Process, please see Loomis (2000, June); Loomis and Hanick (2000b; 2000c); and Reckase, (2000; 2001). For the most part, the same feedback data are provided each time and updated to reflect the most recent round of item ratings. Panelists are carefully instructed in the feedback data to make sure they know the source of the data, the purpose for having the data, how to interpret the data, and how to use the data to adjust their ratings in the next round. At all times, the Achievement Levels Descriptions and concepts/descriptors of borderline performance are pointed to as *the* criteria to use in rating items. Panelists are instructed to take all feedback into consideration—not just one or two pieces. They are instructed that the feedback data are to inform them and that they may decide whether to use it or not.

Panelists are given charts showing their cutscores and the standard deviation of each. The standard deviation is represented as a bar, and panelists are instructed in the fact that the length of the bar is a function of the level of variability of the cutpoints of panelists for that achievement level. Over the rounds, the standard deviations tend to decline, although the cutscores for the groups change only slightly.

Panelists are also given overall student performance data (p-values) for all items in their item rating pool. The percentage of students scoring at each rubric value and the mean score are also reported for constructed response items.

Graphs are provided to show the location of each rater's cutpoint for each of the three levels—three graphs. A secret identification code is used to locate the rater's cutpoint so that this information remains confidential. These Rater Location Charts are a way of presenting interrater consistency data to help panelists determine how consistent their overall ratings—their representations of borderline performance—are relative to other panelists in their group.

Panelists are given holistic data: student performances on a NAEP test booklet scored within 2% points of the cutpoint set in the round of ratings just completed. Given the cutpoints just set by

the ratings of panelists in the group, these booklets represent performance of students at the borderline of each achievement level—at each cutpoint. Panelists are asked to decide whether the performance matches their concept of borderline performance or whether it seems too high or too low. They are instructed in how to change their ratings if either of the latter conclusions is reached. Panelists have the opportunity to review booklets for a total of about 10 students—some at each level, if possible. This direct review of student booklets occurs only after Round 1. They are also given data reporting the percentage of total possible points that students scoring at the cutpoint of each achievement level needed to get correct on the form of NAEP that the panelists were tested in on the first day. Half of the panelists are tested in one test booklet form and the other half in another test booklet form. Panelists are familiar with those items, but the items are not part of their item rating pool. This holistic information helps panelists decide whether their cutpoints seem reasonable, too high, or too low.

Good standard-setting practice generally calls for providing panelists with intrarater consistency data. ACT tried several ways of providing intrarater consistency data, but those methods were not judged to be successful. In fact, ACT eliminated intrarater consistency data from the 1996 Science ALS Process because of concerns that panelist did not really find the data helpful. There is no reason to provide feedback data unless the data actually inform the panelists. If there is doubt, the data should not be provided because there is no good way to judge whether panelists are being informed or misinformed. Further, valuable time may be spent in reviewing data of little or no value to panelists.

Starting with the 1998 NAEP ALS Process, panelists were given Reckase Charts to evaluate their item-by-item ratings. The charts show the probability of correct response for each multiple-choice item for students performing at each point on the score scale. For constructed response items, the charts show the average score for students performing at each point on the score scale. (Loomis, 2000b; Loomis & Hanick, 2000b; 2000c; Loomis & Bourque, 2001; Reckase, 2001). Item ratings are marked for each item, and panelists connect their ratings with colored highlighters to see how consistent their ratings were with respect to student performance and to compare their ratings for each item with their own cutpoint and with the cutpoint for the group as a whole for each level. They can also compare ratings for multiple choice items and constructed response items to evaluate the consistency of their ratings for items of different formats. And, they can compare their ratings for items in different subdomains of the assessment—narrative writing samples, informative writing samples, and so forth. The Reckase Charts were a great breakthrough in feedback data for panelists. Panelists could easily see item characteristic data and understand the relationships among the items and item ratings. In evaluations, Reckase Charts were judged to be the most helpful data.

Another addition to the 1998 NAEP ALS Process was the presentation of consequences data during the rounds of ratings. NAGB had wished to maintain a strictly criterion-referenced standard-setting process, so they did not want to give consequences data to panelists to use in the rating process. Starting in 1994, ACT had provided consequences data to panelists *after* the final round of item ratings were collected. Panelists were told the percentages of students who would score at or above each cutscore set in the final round of ratings, and they were asked to comment on the consequences. In particular, panelists were asked whether they would lower a cutscore in order to increase the percentage of students scoring at or above the level or raise a cutscore in order to decrease the percentage of students scoring at or above the level. Panelists could recommend changes in any, all, or no cutscores. These recommendations were collected and reported to the Governing Board for consideration in reaching their decision on cutscores.

In addition to the consequences questionnaire data just described, ACT collected data in the 1998 field trials to study the impact of consequences data on ratings. In the 1998 studies, one group of panelists was given consequences data as part of the feedback data throughout the process. The other group was not given the data until the last round of item ratings had been collected. The results showed that there was no consistent pattern of differences in cutpoints set by the two groups of panelists. These results led NAGB to approve ACT's request and the recommendation by TACSS to provide consequences data to panelists. Consequences data were first provided to panelists as part of Round 2 feedback. Following Round 3, consequences data were reported for the cutpoints set by each individual panelist in the grade group. Panelists were allowed to give final cutpoints that they wanted to have used to compute the final grade-level cutpoints for reporting to NAGB. Changes in cutpoints recommended by panelists represented only minor changes and adjustments. No panelist took that means as an opportunity to try to sharply alter the grade-level cutpoints. Most panelists seemed pleased with the opportunity to review consequences data and to have the opportunity to make final cutpoint recommendations. They were confident in their ratings and confident in their cutpoints, and they did not choose to make large adjustments despite the very small percentages of students scoring at or above the Advanced achievement level in writing, for example.

Panelists' Evaluations of the Process

Panelists complete evaluations throughout the process. Seven process evaluations are collected from panelists, and these help to monitor the process and document panelists' reactions to each aspect. Data have been collected on some of the same items from all NAEP ALS panelists. Data can be compared across stages in the process (such as rounds of ratings), across grades, and across subjects. The pattern generally found is that panelists are usually more confident in their understanding of the ALDs and in their concept of borderline performance *before* the first round of ratings than they are *during* the first round or two. Confidence and satisfaction tend to increase across rounds of ratings. Nearly all panelists indicate a willingness to sign a statement recommending the achievement levels. The *average* level of confidence, satisfaction, and *positive* responses is generally around 4, which is between *somewhat* (3) confident/satisfied/agree and *totally* (5) confident/satisfied/agree.[8]

General Findings

ACT performs a set of analyses for standard-setting data that are now somewhat standard. Data are evaluated for evidence of statistically significant differences according to the following factors or attributes:

1. Panelist type, race/ethnicity, gender, region
2. Item rating group and table discussion/work group
3. Item format: multiple choice or constructed response (short or extended)

No clear pattern of statistically significant differences has emerged for any of these factors. In some studies for a particular subject, the cutpoints set by teachers are statistically significantly higher than those set by general public panelists at one grade and lower at another. Occasionally, a statistically significant difference will be found between/among panelists according to gender, race/ethnicity, or region, but there is never a consistent pattern across grades.

We do not find statistically significant differences in cutpoints set by panelists in the two item rating pools, and their ratings for common block items are not statistically different. There are occasionally statistically significant differences by table group, but those differences are generally associated with demographic factors of a specific panel member. There is no consistency in this finding and there is no pattern within a subject.

---

[8] Please refer to the sources sited in footnote 7.

We do find a statistically significant difference in the cutpoints that would be set for multiple choice items versus those that would be set for constructed response items. This was a finding in the first pilot study that ACT conducted for the NAEP ALS in 1992, and it has been a persistent finding. Analyses of the 1998 Civics NAEP ALS rating data showed that the differences in cutpoints by item type were small by Round 3.

We examine the relationship between some factors:

1. Interrater consistency with respect to the group cutscore and level of understanding of/confidence in/satisfaction with ALDs, borderlines, rating instructions, and so forth
2. Intrarater consistency and rating changes in subsequent rounds

We have not found a clear pattern of relationship between panelists who have more extreme ratings and their self-reported levels of understanding and confidence in aspects of the process such as the achievement levels descriptions, borderline performance, the tasks, instructions, and so forth. "Outliers" report no more or less confidence in their understanding of the tasks and satisfaction with the outcomes of the process than other panelists. Similarly, we have found no meaningful relationship between interrater consistency and changes in item ratings across rounds. There is no relationship between rater location and the number or magnitude of item rating changes from round to round.

Because the Reckase Charts were very well received by panelists, we attempted to find empirical evidence of the impact on rating changes. Our approach was to consider the conditional p-values associated with the rater's cutscore for a round as the predictor of the item ratings for the subsequent round. Visual inspection of the Reckase Charts had already revealed that no panelist had relied solely on the data in the charts for rating items. Our analyses were to examine whether there was a discernable relationship. There was no clear pattern, however. This finding was taken as evidence that panelists were not relying only on the Reckase Charts as the sole source of feedback and information to guide their item ratings. There was concern that the data in the Reckase Charts would be so easy to use that panelists would focus on the numbers in the charts and disregard the ALDs and borderline performance. These analyses helped to remove those concerns.

We compare process evaluations to those by panelists in previous studies in similar subject areas. For example, we compared responses by civics panelists to those of geography and U.S. history panelists, and we compared 1998 writing panelists' responses to those of 1992 writing panelists. If we find patterns of differences, we dig deeper. In general, we have found that panelists' responses differ somewhat by subject. For example, civics panelists tended to be less positive in their evaluations of the process than writing panelists, although the process was virtually identical for the two groups.

We now expect to find that panelists' responses become more positive after the first round of ratings. Their confidence increases, their understanding of the tasks increases, and their satisfaction with the outcomes increases at each round. One or two disgruntled panelists occasionally emerge from the group. Some panelists have a mission that is not served by training in the process to prepare them to make informed judgments about performance relative to the achievement levels definitions. Those panelists have been relatively few, however. Most panelists appear to be genuinely dedicated to following the procedures and using their best judgment to set performance standards. Those panelists generally reveal that confidence, satisfaction, and understanding increases throughout the rounds of ratings with feedback.

For grade 8 science panelists in 1996, we found that these indicators dropped at the final round. Those panelists had struggled to reach agreement on the meaning of the achievement levels and their concept of borderline performance. Dissatisfaction and a lack of confidence emerged at the end of the process. We reconvened the panelists a few weeks later after more analyses of the data collected from the group. They had the opportunity to evaluate the ALDs, their cutpoints, and their selection of exemplar items. They rather easily reached agreement on minor modifications to each.

Research Studies Related to Validation Issues

ACT has conducted many research studies related to the validity of the NAEP achievement levels. Those can be grouped into three large categories:

1. studies using item mapping procedures as a basis for comparing judgments of performance associated with items to empirical classifications of performance on the items,
2. studies comparing teacher's judgments of performances of their own students to the empirical classifications of those students' performance, and
3. studies comparing judgments of performance represented in test booklets to the empirical classifications of those booklets.

Studies Using Item Mapping Procedures

Studies for the 1992 Mathematics NAEP ALS

ACT used item mapping procedures in 1992 studies with mathematics data, in 1993 with reading data, in 1994 with geography and U.S. history data, and in 1998 with civics data.

In one 1992 study for mathematics, items were classified into achievement levels categories by using a .65 probability of correct response as the criterion (ACT, 1993d). Items were evaluated according to their probability of correct response at the cutpoint of each level. Items were classified at the lowest level for which the .65 probability of correct response criterion was met. Further, the average rating for each item by panelists was used to classify each item. Items were classified at an achievement level for which they had an average rating of .65. The two sets of classifications were compared. For all grades and achievement levels combined, 514 of the 619 items in the 1992 NAEP Math item pool (83%) matched on the judges' rating and response probability. For all grades combined, the match was very high at the Basic level (97%), high at the Proficient level (87%) and high at the Advanced level (72%). Correlation analyses between the two sets of classifications yielded coefficients ranging from a low of .78 for Grade 4 Advanced to a high of .93 for Grade 12 Basic. These data are reported in Table 1.

A second study was conducted with panelists for each of the three grades divided about equally according to experience in the ALS process for mathematics (ACT, 1993d). Six of the 11 panelists had participated in the mathematics ALS and in a content validity study, and five of the panelists were recommended by various stakeholder groups and selected from among persons who had participated in a NAEP Item Anchoring procedure for the 1992 Mathematics NAEP.

Panelists first classified each item in the grade-level item pool independently. They were instructed to classify items into achievement levels that best fit with the ALDs. The categories for classification were Below Basic, Basic, Proficient, Advanced, and Cannot Classify. Then, pairs were formed so that one person had prior ALS experience and one did not. The pairs compared their classifications of items. More than half of the items at each grade level were classified at the same level by both persons in the pairs. Fewer than five items for any grade were classified in non-adjacent levels. This was evidence that persons with and without prior training in the mathematics ALDs made similar judgments about the relationship between the task assessed and

the level of knowledge and skills that a student needed, according to the ALDs, in order to respond correctly.

Most of the items in the 1992 NAEP Mathematics item pool for each grade were classified at the Basic or Below Basic level, and very few were classified at the Advanced level. Those classifications are reported in Table 2. Table 3 shows the average conditional probabilities of correct response for items classified at each level. The finding of the study was that the average probability of correct response for items classified in each achievement level category was at least 60%.[9] That finding was taken as evidence that the achievement levels cutpoints were appropriately matched to the achievement levels descriptions for the mathematics assessment.

The 1994 Reading Revisit Study

The Reading Revisit Study (ACT, 1995a) used the same procedures described in the second mathematics study. This was called a *judgmental item classification* (JIC) procedure, and 8 panelists for each grade participated in this study. In addition, a second classification procedure was used in this study by 32 panelists (10 each for grades 4 and 8, and 12 for grade 12). The second procedure was called the Item Difficulty Categorization (IDC) procedure. The IDC study was designed specifically to answer the question: *Can students do what the ALDs say they should be able to do?* For the IDC, panelists reviewed items that were statistically classified as *can do, can't do,* and *challenging*. Items were placed in the *can do* category if they had a .50 probability of correct response at the lower boundary of the achievement level. Items were placed in the lowest achievement level category for which this criterion was met. The *can't do* category included items with less than a .50 probability of correct response at the upper bound of a particular achievement level category. The remainder of the items were classified in the *challenging* category for the level. The probability of correct response was greater than .50 at some point within the level, but less than .50 at the lower boundary.

Panelists examined items in the categories of each achievement level classified as *can do* or *challenging* to determine whether the knowledge and skills required for the items corresponded to the statements of what students *should* be able to do. Similarly, they examined items in the *can't do* list to determine whether there were any statements in the ALDs describing knowledge and skills that were included in those items. They were asked to compile lists of inconsistencies and share them for discussion with others in their grade group. They were to report specific aspects of the ALDs that lacked support on the basis of the comparisons of items in the IDC. They were then asked to make recommendations for changes based on those.

In all three grade levels, the panelists concluded that students could generally do what the ALDs called for. There were certain exceptions to this, but panelists termed those as *anomalies*, as opposed to patterns. Panelists provided reasons for which items appeared in the *can't do* category despite the fact that they were intended to measure knowledge and skills that students *should* have. For the most part, they found peculiarities in the items that they felt were responsible for the classification. Panelists found items in the *can do* categories for grades 4 and 12 that were not included in the descriptions for achievement at the levels. They recommended that the ALDs be modified to indicate that students at the Basic level can make simple inferences. That change was made.

[9] The *a priori* criterion for the study was set at p=.51, so these findings exceeded the minimal level of acceptable evidence. Further analyses of the data, weighting for the number of items and item ratings, showed the average weighted p-value to be at least 65%.

The results of the JIC procedure provided confirmation that the achievement levels descriptions could be used consistently by panelists to classify assessment items. The pattern of findings showed that panelists' judgments generally match student performances on the items.

Finally, the results of the JIC were compared to the IDC. That is, the judgmental classifications of the items by one group of panelists were compared to the statistical classifications used by the other group of panelists in the study. The "hit" rates for the three levels in each of the three grades were quite high. The results of these analyses are summarized in Table 4. In the JIC procedure, panelists used the achievement levels descriptions to develop the categorizations of items. In the IDC procedure, student performance at each achievement level was used to categorize items. The high correspondence between the two procedures was found to provide compelling evidence that the achievement levels descriptions communicate clearly and accurately with respect to student performance.

The 1999 Civics Item Classification Study[10]

The purpose of the study was to determine whether there was evidence of a reasonable correspondence between the Civics NAEP Achievement Levels Descriptions (ALDs) and the performances of students within achievement levels. Is there evidence that students performing within the cutscore ranges *know and can do* the types of things that the ALDs require for performance within each level?

This small study was conducted at the request of the Achievement Levels Committee of NAGB. A larger study had been planned, but it was canceled after TACSS raised serious concerns about conducting the study. By this time, TACSS was somewhat opposed to studies requiring a response probability when the outcome of the study would be influenced by the p-value used for item mapping. The choice of response probabilities will, in large part, determine the particular match of items to levels. NAGB had not established a response probability to use in the achievement levels-setting (ALS) process, however, so three different probabilities were examined in analyzing the data for this study.

Three teachers at each grade level were recruited from the Iowa City School District. Their task was to use the achievement levels descriptions to classify items in the grade level item pool. The classification forms included a category for Below Basic and Cannot Classify.

In general, the findings of the study showed that there was a reasonable correspondence between the performance of students in each achievement level category and items that represent the knowledge and skills that students *should* have, according to the ALDs.

There was little difference by item type (i.e., multiple choice and constructed response items) in the correspondence of classifications based on teachers' judgments and student performance data.

Twelfth grade teachers reached the highest rate of agreement (98%) in item classifications according to the ALDs and the lowest rate was by eighth grade teachers (71%). Across the three grades, the average rate of agreement was 86%.

Further evaluations were made of the impact of response probabilities on the correspondence between item classifications based on teachers' judgments relative to those based on response probabilities. A 65% response probability showed the highest correspondence with teachers' classifications of items.

---

[10] The complete study is reported in Loomis, 2000a.

Classification data were also evaluated to study the impact of using a correction for guessing on the correspondence between item classifications based on teachers' judgments relative to those based on response probabilities, with and without a correction for guessing. Using a correction for guessing on multiple choice items increased the correspondence with teachers' classifications of items. A 65% probability of correct response, corrected for guessing, is mapped as if the response probability were 74%. A surprisingly large percentage of items in the 1998 Civics NAEP item pool had relatively high guessing parameters. Forty percent of the multiple choice items at grade 4, 54% at grade 8, and 69% at grade 12 had guessing parameters that exceeded the chance probability (25%) of randomly guessing the correct response. The relatively high potential for student guessing would likely lead to a lower correspondence between teachers' judgments of item difficulty and student performances. Teachers did not take guessing into consideration when classifying the items because the ALDs do not address guessing.

In general, the findings from this study were viewed as confirming that the ALDs are interpreted consistently by standard setting panelists and other panelists and that the performance of students within the achievement levels ranges is reasonably well described by the ALDs. That is, it appears that student performance matched the achievement levels descriptions.

Studies Comparing Teachers' Judgments with Empirical Classifications of Student Performance

ACT designed a study for the 1994 NAEP ALS Process in both geography and U.S. history to examine whether there was evidence that achievement levels descriptions were understood by panelists such that panelists would interpret the ALDs consistently from the item rating process to evaluations of their own students (ACT, 1995c). The National Academy of Education (1993) had argued that panelists did not fully understand the achievement levels descriptions and that the results of applying the achievement levels descriptions in the (flawed) item-by-item judgment process would differ from the results of applying the achievement levels descriptions in a holistic judgment process. The study design used by ACT was a modification of a contrasting-groups type design used in one of the studies for the NAE evaluations.[11]

The study design included eighth grade teachers who had participated as panelists in either the pilot study or the operational ALS study. It also included their students in the subject for which they had served as a panel member. If these teachers could use the ALDs to estimate achievement and performances of their students in a way consistent with the students' performance, then this would lend support to the levels set. If not, then it would seem unlikely that people who are not well trained in the ALDs and NAEP procedures would be able to make reasonable interpretations about the meaning of the achievement levels. The study design was seen as offering the "best case scenario." If there were no evidence that the teachers used the achievement levels consistently, then we would assume that most people would misinterpret the outcomes of the NAEP ALS process.

Results from those studies in three different subjects were all similar. Teachers classified their students two times. First, teachers classified each student into one of the four levels (including Below Basic) based on their judgments of the student's overall knowledge and skills in the subject relative to the ALDs. Second, teachers classified each student into one of the four levels based on their judgements of how the student would perform on the NAEP—again, relative to the

[11] A complete description of the ACT study for 1998 was presented at the 2000 Large-Scale Assessment Conference. See Loomis, 2000c. That paper reviewed the NAE study findings and those of the ACT studies for geography and U.S. history.

ALDs. Teachers were aware that their students would be tested (geography and U.S. history) or had been tested (civics) with a special form of NAEP that lasted 100 minutes—twice as long as the actual NAEP in the subject. They were not told which items were included on the test.

Results of these studies indicated that teachers consistently judged the level of subject matter knowledge and skills to be higher than the empirical performance level of the students. Please refer to the data in Tables 5 – 7. They also judged the level of performance that their students were likely to exhibit on the special form of NAEP to be higher than the empirical performance level of the students. Please refer to the data in Tables 8 – 10. Based on these judgements by the teachers, the achievement levels cutscores would appear to have been set too high. Teachers interpreted the achievement levels descriptions holistically, with respect to overall knowledge and skills of their students, to require lower levels of performance than that represented by the cutscores. They interpreted the achievement levels descriptions holistically, with respect to likely performance of their students on NAEP, to require lower levels of performance than that represented by the cutscores.

The differences between the performances of students and judgements of teachers in the first set of classifications were greater than for the second. That is, the more abstract classification of their students (overall knowledge and skills) was less similar to the empirical performance of students than their second classification in which they estimated how students would perform on the assessment.

Those results were evaluated singly and in conjunction with findings from other studies conducted for each of the subjects. Members of the Technical Advisory Committee were neither surprised nor alarmed by the findings. They had anticipated that a "halo effect" would be evident in the judgements of the teachers. To their credit, teachers tend to have high expectations of their students and to expect a high level of performance from them. TACSS had anticipated this directional pattern in the analyses of matches between the classifications based on judgments and those based on student performances. Indeed, these findings were consistent with other studies regarding teachers' abilities to judge student performances on non-classroom test instruments. Other research related to teachers' abilities to judge the performance of their students indicated that teachers are only moderately proficient at estimating the *relative* level of their students' performance on standardized test (Hoge & Coladarci, 1989). That is, teachers can do fairly well in ordering their students' performance on standardized tests. Teachers do less well, however, when it comes to their estimating the *actual* level of their students' achievement relative to the students' standardized test scores. Teachers consistently tend to overestimate their students' standardized test performance (Perry & Meisels, 1996). Relative to the empirical score classifications of student performance on NAEP achievement levels, teachers in these studies reported here also consistently tended to overestimate their students' achievement.

Studies Comparing Judgments of Performance Represented in Test Booklets to the Empirical Classifications of Those Booklets

The booklet classification studies investigated the linkage between the Achievement Levels Descriptions (ALDs) and the cutpoints set to represent student performance with respect to each achievement level. Specifically, the studies examined the extent to which students with scores in the intervals defined by the cutpoints demonstrated knowledge and ability judged to correspond to the ALDs.

The first booklet classification study (BCS) implemented by ACT was designed for the research studies in the 1994 NAEP ALS process (ACT, 1995c). The second BCS was implemented for the 1996 Science ALS process (ACT, 1997b). The third was implemented for the 1998 Civics ALS

process (Loomis, 2000a). The general design of the three studies was the same. There were some differences, however. For the 1994 study with geography and U.S. history, panels were drawn for all three grade levels. Only grades 4 and 8 were investigated in the science study, and only grade 8 was investigated in the civics study. With the exception of the civics study, panelists were recruited according to the same guidelines and criteria used in selecting ALS panelists. For civics, only grade 8 teachers who had participated in either the pilot study or the ALS were included. The civics panelists participated in both the BCS study and the study previously described in which they classified the performances of their students.

Ten panelists participated in each grade group for the geography and U.S. history studies; 13 participated in each grade group for the science study, and there were 11 teachers with complete data used in reporting results for the civics study.

Booklets were selected for the study using the same general criteria. Forty booklets were selected to represent student performance across the four levels marked by the cutpoints for the three NAEP achievement levels. Seven booklets were drawn from within the range of Below Basic performance and seven from the Advanced range. Thirteen booklets were drawn from the range of Basic and 13 from the range of Proficient performance. The level of performance of each student booklet was determined by using NAEP plausible values for the geography, U.S. history, and science studies. Student booklets for the civics BCS were those that had been administered in the special study including enough items from the civics NAEP to produce a reliable estimate of student performance without the use of plausible values. The actual examination booklets for students having 4 of the 5 plausible values within a specific cutscore range were acquired for the earlier studies.

The criteria for selecting booklets for the civics BCS were somewhat different. TACSS recommended that the booklets be selected so that they were neither right around the cutscores nor all clustered at the midpoint of the achievement levels ranges. This was to eliminate booklets with scores that would likely be ambiguous with respect to the empirical scores for classification. Of the 461 booklets that met these criteria, booklets scored within the achievement levels ranges were randomly selected. Some substitutions were made to maximize the number of different schools represented at each level of achievement in the study. Some substitutions were also made to balance the number of booklets in each form at each level.

In each BCS, panelists were asked to evaluate each booklet to determine whether the performance exhibited was consistent with the definition for Basic, Proficient, or Advanced. Panelists were given scores for neither booklets nor items. Booklets were distributed in random order. Panelists recorded their classification of each booklet on a form.

The findings were consistent for the studies in the four subjects. Panelists tended to classify the performance exhibited in the student test booklets at the empirical level or on level lower. Very few classifications were at a level higher than the empirical classification level. The average overall hit rates are reported in Table 11. The classifications by empirical score level and judgement level are presented for geography in Tables 12 – 14, for U.S. history in Tables 15 – 17, for science in Tables 18 – 19, and for civics in Table 20.

Analyses of data by panelist type in the earlier years showed no statistically significant differences. It was somewhat surprising to find that the teachers were no more accurate than other types of panelists in judging the performance, relative to the achievement levels definitions, represented in the booklets.

Based on the evidence from the booklet classification studies, one would conclude that the cutscores in each subject were set too low. That is, performance would need to be higher in the booklets classified empirically in each achievement level category in order to increase the match between the empirical score classifications and judgments of the performance.

The BCS evidence was not the only evidence for judging the cutscores and the ability of panelists to interpret the ALDs in a manner consistent with that represented by the cutscores. Evidence from the BCS for geography and U.S. history was paired with findings from the studies in which teachers had classified their own students. The findings from the two sets of studies for these two subjects pointed to different conclusions. The cutscores were set lower than the results of the BCS would suggest and higher than the results of the teacher classifications would suggest.

The 1998 civics study design, as noted above, was different in that it included only teachers and the booklets in the BCS represented individual student performance scores. It was possible to match the classifications of students by their teachers and the classifications of the same students' booklets by the teachers. Those results showed that the teachers classified the booklets lower than they had classified the expected performance of the students, and expected performance was classified lower than the overall level of student knowledge and skill in civics. Please see Table 21.

Having the same civics panelists participate in both parts of the studies that had been conducted separately with different sets of panelists for geography and U.S. history provided a control on panelists to make direct comparisons across different classification tasks. The findings from the study with the civics panelists were the same as those for the two sets of panelists in the two studies each for geography and U.S. history. Although the findings seemed logical and somewhat likely, it was not clear whether the differences in findings for the BCS versus the classifications of students were a result of differences in panelists' judgments. The design of the civics study helped to clarify this. The conclusion is that teachers (perhaps all people, but the evidence does not extend beyond teachers) will judge the performance of their own students to be somewhat higher than the students' performance when the identity of the student is known and when the basis of the judgment is somewhat subjective. When teachers (perhaps all people, but the evidence does not extend beyond teachers) judge the performance of their own students in a "blind test" (student identify is not known), they tend to judge the performance to be somewhat lower than the students' performance on the assessment.

What does this say about the validity of the NAEP achievement levels and the probability of consistent interpretations of achievement levels descriptions? The findings from these studies showed the same or slightly higher levels of correspondence ("hit rates") as those reported in other studies for which teachers classify student performances on other assessments. Teachers (perhaps all people) judge the rank ordering of student performance on assessments rather accurately. They do not classify the students in levels with the same level of accuracy. The tendency is to be more lenient when the classification is more abstract and subjective and to be harsher when the classification is more concrete and objective. Further, it appears that anonymous performance may be judged more severely than performance of one's own students.

Conclusions

In the end, there is no way to know with certainty that cutscores are valid. Collecting validity evidence about standard setting is difficult. Standards are based on judgments. There is no true standard against which to judge the outcome of a standard-setting process (Kane, 1994; Cizek, 2001; Zieky, 2001). In many ways, the NAEP achievement levels serve as a sort of "gold standard" (Kane, 1994) against which state standards are judged. So, the NAEP achievement

levels serve as a standard against which other standards may be judged, but there is no standard against which to judge NAEP achievement levels.

Congress has stated that the achievement levels must be shown to be useful, reasonable, and valid. NAGB asks ACT for evidence of the validity of the achievement levels. They seek additional data and findings to inform their decisions regarding where to set the cutscores. Finding validity evidence for NAEP poses even greater problems than is the case for most other assessments. There are no individual student scores, and no student is tested over enough items to provide an adequate measure of knowledge and skills to produce a valid measure at the individual student level.[12] There is no student performance in NAEP to compare to performance on another assessment. Even when individual-level student data were collected in the special studies conducted by ACT, there were no absolute criteria by which to evaluate the findings. The findings seemed reasonable; there was no evidence that the outcomes of the ALS process were not valid. But, was the evidence of validity sufficient? That is a judgment.

Further, the conditions for NAEP administration are quite different from those of other assessments, making comparisons generally inappropriate. For example, because NAEP is a survey of students at specific grade levels, it is quite possible that the students tested have not recently taken a course in the subject. It is difficult to judge how much to "discount" student performance on NAEP when comparisons are made to performance on other assessments that are tied to learning goals of the school, district, state, or program. The Governing Board typically considers performance on the Advanced Placement (AP) Test, if there is one in the subject, as a guide to the reasonableness of the Advanced level cutscore for grade 12. They have special computations run to produce a measure of the percentage of the grade 12 cohort scoring 3 or higher on the test. NAGB members understand the ways in which the AP differs from NAEP and why performance on the two is not directly comparable, even in the aggregate. This provides *some* indication for one of the nine cutscore. But, whether the grade 12 Advanced achievement level is reasonable, relative to AP performance in the subject, is a matter of judgment. How close do the percentages need to be to judge the NAEP cutscore for the grade 12 Advanced level as *reasonable*? How much should *reasonable* count toward *valid*? How much weight should be assigned to data for subjects with relatively few AP scores? How should judgments be made for subjects for which there is no AP test?

NAGB also requests ACT to provide contextual or corroborative data to help them reach their decision regarding the reasonableness of the recommendations put before them. When possible, ACT provides comparison data from subjects for which NAEP achievement levels have already been set. The comparability of subjects and the relative level of performance expected in each is a matter of judgment. ACT typically compiles data on course-taking patterns in the subject for students at or near the grades surveyed by NAEP. ACT tries to provide a context for NAGB to use in judging the cutscores and performance distributions.

ACT also develops charts and graphs from NAEP student background data and teacher and school questionnaire data to help the NAGB policy makers understand more about the students who were included in the NAEP survey. Some of these data are helpful, and some only add to the confusion. Motivation data, for example, present a counterintuitive picture at first glance. Students who report that they tried much harder to perform well on NAEP than they usually try

---

[12] Getting approval for the collection of student performance data for the first validation study in 1995 that involved student performance judgements by their own teachers required months and months of effort and involvement of persons at high levels within the U.S. Department of Education. At that time, there was genuine concern about producing individual scores in the NAEP program.

on other assessments do not score as high as students who report that their effort was about the same, or less, as for other assessments. Students who have recently taken or are taking courses in the subject may be found to score lower than other students. The data available for judging the reasonableness of NAEP performance are not straightforward indicators. Judgments are required.

ACT and TACSS have given considerable thought to the question of how to provide convincing evidence regarding the validity of the achievement levels. We believe that data collected in the course of the NAEP ALS process provide confirmatory evidence. Panelists understand the achievement levels and they can apply the ALDs consistently in contexts outside standard setting. The evidence seems to suggest that the cutscores are about right—neither as high as teachers might set them, based on actual student performance, nor as low as they would probably place them when judging their own students in a somewhat subjective context.

Procedural validity is the ultimate evidence that we have to offer. To some extent, that has come to be expected—to be taken as a *given* of the NAEP ALS process. While NAGB, as a policy board, appreciates the essential role of procedural validity, they still face the charge to show that the levels are useful, reasonable, and valid. NAGB makes every effort to find something other than the popular opinion poll for making their judgments. As is the case for the judgments made by standard-setting panelists, NAGB's judgments must be informed. It is a struggle.

REFERENCES

ACT. (1993a). *Setting achievement levels on the 1992 NAEP in reading: Final report.* Iowa City, IA: Author.

ACT. (1993b). *Setting achievement levels on the 1992 NAEP in mathematics: Final report.* Iowa City, IA: Author.

ACT. (1993c). *Setting achievement levels on the 1992 NAEP in writing: Final report.* Iowa City, IA: Author.

ACT. (1993d). *Setting achievement levels on the 1992 NAEP in reading, mathematics, and writing: A technical report on reliability and validity.* Iowa City, IA: Author.

ACT. (1994a). *Results of the 1994 geography NAEP achievement levels setting pilot study.* Iowa City, IA: Author.

ACT. (1994b). *Results of the 1994 U.S. history National Assessment of Educational Progress achievement levels-setting pilot study.* Iowa City, IA: Author.

ACT. (1995a). *NAEP reading revisit: An evaluation of the 1992 achievement levels descriptions.* Iowa City, IA: Author.

ACT. (1995b). *Preliminary report on the 1994 NAEP achievement levels setting process for U.S. history and geography.* Iowa City, IA: Author.

ACT. (1995c). *Research studies on the achievement levels set for the 1994 NAEP in geography and U.S. history.* Iowa City, IA: Author.

ACT. (1997a). *Setting achievement levels on the 1996 NAEP in science: Final report, Volume III: Achievement levels-setting study.* Iowa City, IA: Author.

ACT. (1997b). *Setting achievement levels on the 1996 NAEP in science: Final report, Volume IV: Validity evidence and special studies.* Iowa City, IA: Author.

Bay, L. (1998). *Booklet classification as a standard setting method for the 1998 NAEP writing: the issue of booklets to be classified.* A draft report prepared for the meeting of TACSS, Chicago, IL.

Bay, L. (2000). Setting achievement levels on the 1998 National Assessment of Educational Progress in writing: Performance profiles. In Loomis, S.C. (Ed.), *Developing achievement levels on the 1998 National Assessment of Educational Progress in Writing: Research studies*. Iowa City, IA: ACT.

Cizek, G. J. (1993). *Reactions to National Academy of Education report, Setting performance standards for student achievement*. Washington, DC: National Assessment Governing Board.

Cizek, G.J. (1996). Standard-Setting Guidelines. *Educational Measurement: Issues and Practice*, 15(1), 13-21.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hanson, B.A. (1998). *Calculating writing cutpoints for booklet classification in the second field trial*. A paper presented to the TACSS, Chicago, IL.

Hanson, B.A., Bay, L. & Loomis, S.C. (1998). *Booklet classification study*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Hoge, R. D., & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Review of Educational Research*, 59, 297-313.

Impara, J.C. and Plake, B.S. (1996). *Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method*. Paper presented at the annual meeting of the National council of Measurement in Education, New York.

Kane, M.T. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461.

Kane, M.T. (1995). Examinee-centered vs. task-centered standard setting. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (119-139). Washington, DC: National Assessment Governing Board and national Center for Education Statistics.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in standard setting. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kolstad, A. (1996, April). *The response probability convention embedded in reporting prose literacy levels from the 1992 National Adult Literacy Survey*. Paper presented at the annual meeting of the American Education Resort Association, Chicago.

Kolstad, A. Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Report prepared for the National Center for Education Statistics, Washington, DC.

Loomis, S.C. (2000a). *Developing achievement levels on the 1998 National Assessment of Educational Progress in civics; Validation research*. Iowa City, IA: ACT.

Loomis, S.C. (2000b, April). *Feedback in the NAEP Achievement Levels Setting Process*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Loomis, S.C. (2000c, June). *Research on the use of holistic methods for the NAEP achievement levels-setting process*. Paper presented at the CCSSO Large-Scale Assessment Conference. Snowbird, UT.

Loomis, S.C. (2000d, April). *Research study of the 1998 civics NAEP achievement levels*. Paper presented at the annual meeting of the American Education Research Association, New Orleans.

Loomis, S.C., Bay, L., Yang, W.L., & Hanick, P.L. (1999, April). *Field trials to determine which rating method(s) to use in the 1998 NAEP achievement levels-setting process for civics and writing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Loomis, S.C., & Hanick, P.L. (2000a). *Setting achievement levels on the 1998 National Assessment of Educational Progress in civics and writing: Finalizing the achievement levels descriptions*. Iowa City, IA: ACT.

Loomis, S.C., & Hanick, P.L. (2000b). *Setting achievement levels on the 1998 National Assessment of Educational Progress in civics: Final report*. Iowa City, IA: ACT.

Loomis, S.C., & Hanick, P.L. (2000c). *Setting achievement levels on the 1998 National Assessment of Educational Progress in writing: Final report*. Iowa City, IA: ACT.

Loomis, S.C., Hanick, P.L., Bay, L., & Crouse, J.D. (2000a). *Setting achievement levels on the 1998 National Assessment of Educational Progress in civics interim report: Field trials*. Iowa City, IA: ACT.

Loomis, S.C., Hanick, P.L., Bay, L., & Crouse, J.D. (2000b). *Setting achievement levels on the 1998 National Assessment of Educational Progress in writing interim report: Field trials*. Iowa City, IA: ACT.

Loomis, S.C., Hanick, P.L., & Yang, W.L. (2000a). *Setting achievement levels on the 1998 National Assessment of Educational Progress in civics interim report: Pilot study*. Iowa City, IA: ACT.

Loomis, S.C., Hanick, P.L., & Yang, W.L. (2000b). *Setting achievement levels on the 1998 National Assessment of Educational Progress in writing interim report: Pilot study*. Iowa City, IA: ACT.

Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. *In Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (221-263). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Mehrens, W.A. & Cizek, G.J. (2001). Standard setting and the public good: Benefits accrued and anticipated. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

National Academy of Education (1993). *Setting performance standards for student achievement*. Stanford, CA: Author.

National Assessment Governing Board (1990). *Setting appropriate achievement levels for the National Assessment of Educational Progress: Policy framework and technical procedures*. Washington, DC: Author.

National Research Council (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell (Eds.). Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment. Washington, DC: National Academy Press.

Perry, N.E. & Meisels, S.J. (1996). *How Accurate Are Teacher Judgments of Students' Academic Performance?* (Working paper No. 96-08). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Public Law 100-297 (1988). National assessment of educational progress improvement act (Article No. USC 1221). Washington, DC.

Reckase, M.D. (1998). *Setting Standards to be consistent with an IRT item calibration.* Iowa City, IA: ACT.

Reckase, M.D. (2000). *The evolution of the NAEP achievement levels-setting process: A summary of the research and development efforts.* Iowa City, IA: ACT.

Reckase, M.D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Reckase, M.D. & Bay, L. (1999, April). *Comparing two methods for collecting test-based judg*ments. Paper presented at the annual meeting of the National Council on Measurement in Education, 1999, Montreal.

Zieky, M.J. (2001). So much remains the same: Conception and status of validation in setting standards. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zwick, R., Senturk, D., Wang, J. & Loomis, S.C. (2000). *An investigation of alternative methods for item mapping in the National Assessment of Educational Progress.* Iowa City, IA: ACT.

Table 1

Mathematics NAEP Item Classification Study

Comparing Panelists' Average Item Ratings to Response Probabilities at the Cutpoint

	Observed p-values ≥	Expected p-values ≥	Percent of Items Meeting Criterion	Pearson Product-Moment Correlations
Grade 4				
Basic	9	9	100%	.87
Proficient	85	75	88	.90
Advanced	94	74	78	.78
Total	188	157	84	.85
Grade 8				
Basic	33	32	97%	.92
Proficient	123	106	86	.92
Advanced	62	34	59	.85
Total	218	172	79	.90
Grade 12				
Basic	34	33	97%	.93
Proficient	110	98	89	.92
Advanced	69	54	78	.83
Total	213	185	87	.89
All Grades Combined				
Basic	76	74	97%	.91
Proficient	318	278	87	.91
Advanced	225	162	72	.82
Total	619	514	83	.88

Table 2

Mathematics Item Classification Panel Study:

Number of Items Classified at Each Achievement Level, by Grade

Grade	< Basic	Basic	Proficient	Advanced
4^{th}	3	88	52	12
8^{th}	26	68	75	13
12^{th}	22	82	64	11

Table 3

Estimated Average Percentage of Students with Scores with Achievement Levels Who Correctly Answered Items Classified at the Corresponding Achievement Level

	< Basic	Basic	Proficient	Advanced
4th Grade	65.1%	59.0%	62.5%	65.7%
# Items	3	88	52	12
8th Grade	65.9	69.8	61.1	68.7
# Items	26	68	75	13
12th Grade	63,7	62.7	63.7	74.2
# Items	22	82	64	11

Table 4

Average Correspondence Between Item Difficulty Categorizations at Each Achievement Level and Judgmental Item Categorizations for Items

Grade Level	Basic	Proficient	Advanced
4th Grade	48	79	93
8th Grade	80	81	92
12th Grade	66	90	97

Table 5

Percentage of Students Classified within Achievement Level Categories Based on Overall Geography Knowledge and Skills Relative to the Empirical (MLE) Score Classifications

Achievement Level Classification of Overall Geography Knowledge and Skills	Achievement Level Classification of MLE Score Estimates of Student Performance				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	3.40 (3.91)	2.89 (3.82)	0.33 (0.83)	0.0 (0.0)	6.63 (6.27)
Basic	8.74 (9.15)	18.91 (10.55)	5.49 (6.01)	0.39 (1.01)	33.52 (14.84)
Proficient	3.31 (7.26)	18.50 (9.70)	18.22 (13.34)	1.20 (1.98)	41.22 (18.02)
Advanced	0.15 (0.48)	3.09 (3.25)	11.49 (5.57)	3.91 (4.10)	18.64 (6.74)
Total	15.60 (16.60)	43.38 (10.95)	35.53 (16.82)	5.49 (5.41)	P_A=.44 P_E=.31 K =.19

Table 6

Percentage of Students Classified within Achievement Level Categories
Based on Overall U.S. History Knowledge and Skills
Relative to the Empirical (MLE) Score Classifications

Achievement Level Classification of Overall U.S. History Knowledge and Skills	Achievement Level Classification of MLE Score Estimates of Student Performance				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	5.80 (4.67)	6.05 (7.14)	0.0 (0.0)	0.0 (0.0)	11.85 (9.15)
Basic	5.70 (6.01)	20.90 (12.97)	4.60 (6.93)	0.18 (0.72)	32.50 (14.18)
Proficient	3.77 (5.77)	22.16 (10.90)	14.92 (13.45)	1.26 (3.35)	42.11 (15.40)
Advanced	0.44 (1.48)	3.04 (5.67)	8.71 (6.84)	1.36 (1.93)	12.38 (11.00)
Total	15.71 (12.46)	53.27 (14.40)	28.23 (15.99)	2.79 (4.30)	P_A=.48 P_E=.31 K =.25

Table 7

Percentage of Students Classified within Achievement Level Categories
Based on Overall Civics Knowledge and Skills
Relative to the Empirical (MLE) Score Classifications

Table N = 414	Achievement Level Classification of Overall Civics Knowledge and Skills (SCS#1)			
Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	Below Basic (n=72)	Basic (n=118)	Proficient (n=119)	Advanced (n=105)
Below Basic (<149.2) (n=64)	**8.7%** (n=36)	5.6% (n=23)	1.2% (n=5)	0.0 (n=0)
Basic (149.2 – 165.39) (n=189)	8.2 (n=34)	**19.1** (n=79)	15.0 (n=62)	3.4 (n=14)
Proficient (165.4 – 177.89) (n=140)	0.5 (n=2)	3.9 (n=16)	**11.8** (n=49)	17.6 (n=73)
Advanced (≥ 177.9) (n=21)	0.0 (n=0)	0.0 (n=0)	0.7 (n=3)	**4.4** (n=18)

Bold entries are for cells that would represent "hits" or agreement.

P_A=.440
P_E =.267
K = .243

Table 8

Percentage of Students Classified within Achievement Level Categories Based on Expected Performance on the Special Form of the Geography NAEP Relative to the Empirical (MLE) Score Classifications

Judgment of Achievement Level Classification of Expected Performance	Achievement Level Classification of MLE Score Estimates of Student Performance				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	4.65 (5.06)	4.67 (4.52)	0.61 (1.24)	0.02 (0.88)	10.13 (7.10)
Basic	8.64 (9.78)	18.37 (10.76)	6.19 (6.08)	0.18 (0.56)	33.38 (14.59)
Proficient	2.19 (4.60)	17.09 (8.99)	17.59 (12.35)	1.31 (2.21)	38.19 (16.36)
Advanced	0.11 (0.49)	3.25 (2.98)	11.14 (5.09)	3.80 (4.05)	18.30 (7.23)
Total	15.59 (16.60)	43.38 (10.95)	35.53 (16.82)	5.49 (5.41)	P_A=.44 P_E=.31 K =.19

Table 9

Percentage of Students Classified within Achievement Level Categories Based on Expected Performance on the Special Form of the U.S. History NAEP Relative to the Empirical (MLE) Score Classifications

Judgment of Achievement Level Classification of Expected Performance	Achievement Level Classification of MLE Score Estimates of Student Performance				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	6.89 (5.14)	7.59 (7.29)	0.40 (1.11)	0.0 (0.0)	14.88 (9.62)
Basic	5.20 (5.42)	25.22 (12.12)	7.95 (14.96)	0.18 (0.72)	33.60 (15.19)
Proficient	3.25 (5.63)	20.13 (11.57)	15.04 (12.95)	1.19 (2.56)	39.62 (16.33)
Advanced	0.37 (1.47)	2.17 (4.13)	7.96 (5.17)	1.42 (2.00)	11.91 (8.48)
Total	15.71 (12.46)	53.27 (14.40)	28.23 (15.99)	2.79 (4.30)	P_A=.49 P_E=.32 K =.25

Table 10

Percentage of Students Classified within Achievement Level Categories Based on Expected Performance on the Special Form of the Civics NAEP Relative to the Empirical (MLE) Score Classifications

Table N = 414	Achievement Level Classification of Expected Student Performance on the Special Form of the Civics NAEP (SCS#2)			
Achievement Level Classification of MLE Score Estimates of Student Performance (ACT NAEP-Like Cutscores)	Below Basic (n=87)	Basic (n=115)	Proficient (n=119)	Advanced (n=93)
Below Basic (<149.2) (n=64)	**9.7%** (n=40)	5.3% (n=22)	0.5% (n=2)	0.0 (n=0)
Basic (149.2 – 165.39) (n=189)	10.9 (n=45)	**17.9** (n=74)	15.0 (n=62)	1.9 (n=8)
Proficient (165.4 – 177.89) (n=140)	0.5 (n=2)	4.6 (n=19)	**12.6** **(n=52)**	16.2 (n=67)
Advanced (≥ 177.9) (n=21)	0.0 (n=0)	0.0 (n=0)	0.7 (n=3)	**4.4** (n=18)

Bold entries are for cells that would represent "hits" or agreement.

P_A=.446
P_E =.268
K = .243

Table 11

Average "Hit Rate" for Booklet Classification Studies, by Grade

Grade	Geography	U.S. History	Science	Civics
4	43%	66%	49%	NA
8	71	51	56	56
12	46	71	NA	NA

Table 12
Average Percent Agreement of Empirical and
Judgmental Booklet Classifications of Booklets
Geography Grade 4

Judgmental	Empirical				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.25 (0.79)	20.75 (10.07)	5.25 (4.92)	0.00 (0.00)	43.25 (13.95)
Basic	0.25 (0.79)	11.50 (9.87)	16.75 (3.74)	2.75 (3.22)	31.25 (10.82)
Proficient	0.00 (0.00)	0.25 (0.79)	9.50 (5.37)	10.25 (2.75)	20.00 (7.82)
Advanced	0.00 (0.00)	0.00 (0.00)	1.00 (1.75)	4.50 (2.30)	5.50 (3.50)
Total	17.50	32.50	32.50	17.50	P_A=.43 P_E=.25 K =.23

Table 13
Average Percent Agreement of Empirical and
Judgmental Booklet Classifications of Booklets
Geography Grade 8

Judgmental	Empirical				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	15.95 (1.99)	4.60 (4.60)	0.00 (0.00)	0.54 (1.14)	22.08 (6.84)
Basic	2.97 (1.99)	25.95 (6.14)	8.92 (5.26)	0.00 (0.00)	37.84 (9.70)
Proficient	0.0 (0.0)	1.62 (2.91)	18.11 (5.10)	4.32 (2.91)	24.05 (6.17)
Advanced	0.0 (0.0)	0.27 (0.85)	5.41 (3.60)	11.35 (3.07)	17.03 (5.85)
Total	18.92	32.43	32.43	16.22	P_A=.71 P_E=.27 K =.60

Table 14
Average Percent Agreement of Empirical and
Judgmental Booklet Classifications of Booklets
Geography Grade 12

Judgmental	Empirical				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.00 (1.05)	18.00 (10.26)	6.75 (8.42)	0.00 (0.00)	41.75 (18.30)
Basic	0.50 (1.05)	16.25 (9.37)	20.50 (6.75)	1.00 (2.42)	38.25 (13.39)
Proficient	0.00 (0.00)	0.75 (1.21)	6.50 (4.89)	5.50 (4.68)	12.75 (3.81)
Advanced	0.00 (0.00)	0.00 (0.00)	1.25 (3.17)	6.00 (5.43)	7.25 (7.59)
Total	17.50	35.00	35.00	12.50	P_A=.46 P_E=.26 K =.27

Table 15
Average Percent Agreement of Empirical and
Judgmental Booklet Classifications of Booklets
U.S. History Grade 4

Judgmental	Empirical				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.94 (2.17)	7.35 (3.73)	0.00 (0.00)	0.00 (0.00)	25.29 (5.58)
Basic	2.65 (2.17)	21.47 (4.81)	12.06 (7.78)	0.29 (0.93)	36.47 (8.90)
Proficient	0.00 (0.00)	3.53 (3.54)	18.53 6.80)	5.88 (2.40)	27.94 (7.50)
Advanced	0.00 (0.00)	0.00 (0.00)	1.76 (2.06)	8.53 (2.57)	10.29 (4.22)
Total	20.59	32.35	32.35	14.71	P_A=.66 P_E=.28 K =.54

Table 16
Average Percent Agreement of Empirical and
Judgmental Booklet Classifications of Booklets
U.S. History Grade 8

Judgmental	Empirical				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.95 (0.00)	18.72 (2.97)	2.31 (1.89)	0.00 (0.00)	38.97 (4.15)
Basic	0.00 (0.00)	11.79 (3.24)	16.15 (5.68)	0.77 (1.24)	28.72 (6.71)
Proficient	0.00 (0.00)	0.26 (0.81)	12.31 (5.64)	8.72 (2.76)	21.28 (5.80)
Advanced	0.00 (0.00)	0.00 (0.00)	2.56 (1.21)	8.46 (3.43)	11.03 (4.02)
Total	17.95	30.77	33.33	17.95	P_A=.51 P_E=.25 K =.34

Table 17
Average Percent Agreement of Empirical and
Judgmental Booklet Classifications of Booklets
U.S. History Grade 12

Judgmental	Empirical				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	21.61 (1.56)	7.10 (6.94)	0.65 (1.36)	0.00 (0.00)	29.36 (8.11)
Basic	0.97 (1.56)	22.90 (7.04)	8.71 (6.81)	0.00 (0.00)	32.58 (9.79)
Proficient	0.00 (0.00)	2.26 (2.18)	14.84 (8.08)	4.84 (5.74)	21.94 (9.83)
Advanced	0.00 (0.00)	0.00 (0.00)	4.84 (3.80)	11.29 (5.74)	16.13 (7.29)
Total	22.58	32.26	29.03	16.13	P_A=.71 P_E=.26 K =.60

Table 18
Average Percent Agreement of Judgmental Classifications
and Empirical Classifications of Booklets
Science Grade 4

Empirical	Judgmental				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.6 (0.9)	0.2 (0.7)	0.2 (0.7)	0.0 (0.0)	17.9
Basic	17.4 (10.2)	15.6 (10.3)	0.4 (0.9)	0.0 (0.0)	33.3
Proficient	5.7 (9.8)	23.1 (6.1)	15.8 (5.9)	1.6 (2.4)	46.2
Advanced	0.0 (0.0)	0.4 (0.9)	1.8 (1.2)	0.4 (0.9)	2.6
Total	40.6 (18.8)	39.3 (13.3)	18.1 (6.2)	2.0 (3.2)	P_A=.49 P_E=.29 K=.29

Table 19
Average Percent Agreement of Judgmental Classifications
and Empirical Classifications of Booklets
Science Grade 8

Empirical Classification	Judgmental Classification				
	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	17.4 (1.1)	0.6 (1.1)	0.0 (0.0)	0.0 (0.0)	17.9
Basic	13.4 (8.7)	18.3 (8.3)	1.6 (2.8)	0.0 (0.0)	33.3
Proficient	1.0 (1.2)	19.5 (9.2)	18.3 (8.8)	4.7 (6.3)	43.6
Advanced	0.0 (0.0)	0.4 (0.9)	3.2 (2.0)	1.6 (2.1)	5.1
Total	31.8 (10.5)	38.9 (10.7)	23.1 (10.0)	6.3 (8.1)	P_A=.56 P_E=.29 K=.37

Table 20

Average Percent Agreement of Grade 8 Civics Teachers' Classifications of Student Booklets into Achievement Level Categories and Empirical Score Classifications of Student Booklets into Achievement Level Categories

Table N =440	Achievement level classification of student booklets by teachers			
Achievement level classification by empirical scores of student booklets (ACT NAEP-Like Cutscores)	Below Basic (n=137)	Basic (n=155)	Proficient (n=104)	Advanced (n=44)
Below Basic (<149.2) (n=77)	**15.9%** (n=71)	1.4% (n=6)	0.0 (n=0)	0.0 (n=0)
Basic (149.2 – 165.39) (n=143)	14.3 (n=63)	**17.7** (n=78)	0.5 (n=2)	0.0 (n=0)
Proficient (165.4 – 177.89) (n=143)	0.6 (n=3)	16.2 (n=71)	**14.1** (n=62)	1.6 (n=7)
Advanced (≥ 177.9) (n=77)	0.0 (n=0)	0.0 (n=0)	9.1 (n=40)	**8.4** (n=37)

Bold entries are for cells that would represent "hits."

P_A=.561
P_E =.263
K = *.404*

Table 21

Relationships Between and Among Empirical Score Performance Classifications of Grade 8 Students in the Special Civics NAEP Study and Their Teacher's Judgment Classifications of Performance by NAEP Achievement Levels

		Overall Knowledge & Skills Level				Expected NAEP Performance Level				Booklet Classification			
		BB	B	P	A	BB	B	P	A	BB	B	P	A
Empirical Classification	BB (6)	3	3	0	0	5	1	0	0	5	1	0	0
	B (12)	1	6	5	0	1	6	5	0	7	4	1	0
	P (12)	0	2	4	6	0	3	6	3	0	7	5	0
	A (7)	0	0	1	6	0	0	2	5	0	0	6	1
	Total	4	11	10	12	6	10	13	8	12	12	12	1
Overall Knowledge & Skills	BB (4)					4	0	0	0	5	1	0	0
	B (11)					2	9	0	0	7	4	1	0
	P (10)					0	1	9	0	0	7	5	0
	A (12)					0	0	4	8	0	0	6	1
	Total					6	10	13	8	12	12	12	1
Expected Performance	BB (6)									5	1	0	0
	B (10)									5	4	1	0
	P (13)									2	5	6	0
	A (8)									0	2	5	1
	Total									12	12	12	1